专题导读：深度学习（Deep Learning）是机器学习的分支，是一种以人工神经网络为架构，对数据进行表征学习的方法。它能够在发现数据的分布式特征表示的同时实现人们对于复杂事务处理的自动化要求。自深度学习出现以来，它已成为很多领域，尤其是在计算机视觉和语音识别中，成为各种领先系统的一部分。农业领域深度学习的应用主要集中于作物表型分析、植物养分含量估计、病虫草害识别、基于遥感图像解译的植物识别、土壤覆盖分类、动物行为识别和分类，以及生物信息分析等领域，且分析结果比传统算法有很大提升。可以预见，深度学习技术将会在农业领域加速渗透，大放异彩。

深度学习在植物基因组学与作物育种中的应用现状与展望

侯祥英[1]，崔运鹏[2*]，刘　娟[2]

（1. 淄博市农业科学研究院，淄博　255020；

2. 农业农村部农业大数据重点实验室，中国农业科学院农业信息研究所，北京　100081）

摘　要：[目的／意义]随着单细胞测序、高通量技术的突破，植物基因组学也取得了巨大进步，可以低成本获取多维全基因组分子表型的海量数据。深度学习技术可以作为强大的数据挖掘工具对获取的分子表型进行进一步预测和解释。当前研究表明，深度学习在植物基因组学与作物育种研究任务中取得显著效果。但目前尚缺乏对于深度学习在植物基因组学中应用的完整综述。[方法／过程]本文首先概述了深度学习方法背景，包括最新的图神经网络；随后着重从基因特性、蛋白质特性方面综述了基因组学和深度学习交叉领域的两个突出问题：1）如何对从植物基因组 DNA 序列到分子表型的信息流进行建模？2）如何使用深度学习模型识别自然种群中的功能变异？[结果／结论]本文总结了当前研究中如何应用传统深度学习算法、图深度学习、生成对抗网络以及可解释性 AI 等方法解决上述两个问题。最后分析了深度学习在未来植物基因组学研究和作物遗传改良中的发展前景。

关键词：植物基因组学；作物育种；深度学习；图深度学习；综述

中图分类号：S-1；Q-31；TP399　　**文献标识码：**A　　**文章编号：**1002-1248（2022）08-0004-15

引用本文：侯祥英，崔运鹏，刘娟. 深度学习在植物基因组学与作物育种中的应用现状与展望[J]. 农业图书情报学报，2022，34（8）：4-18.

收稿日期：2022-02-28

基金项目：中国农业科学院院增项目“作物育种深度分析技术”（2020ZLK005）；国家社科基金重大项目“中国古农书的搜集、整理与研究”（21&ZD332）

作者简介：侯祥英（1971-），女，农艺师，淄博市农业科学研究院，研究方向为经济作物栽培与育种。刘娟（1978-），女，副研究员，研究方向为农业大数据应用研究，数据治理

*通信作者：崔运鹏，研究员，中国农业科学院农业信息研究所农业大数据挖掘研究室，主任。Email：cuiyunpeng@caas.cn

1 引 言

植物基因组学分析与育种的研究目标是对植物全生命周期的信息流进行研究。该信息流从基因组 DNA 序列分析开始，并在植物表型研究或作物物种、农艺性状等方面的研究结束。介于基因层面和植物表型层面信息之间的是通过转录和翻译传递的信息流，这就是弗朗西斯·克里克（FRANCIS）提出的“分子生物学中心法则”[1]。中心法则中的每一步都不仅可以看作是传递，还可以看作是前一步遗传信息的转化。所涉及的分子特征统称为“分子表型”，以将它们与终端特征区分开来。随着单细胞测序、高通量测序技术的突破，植物基因组学、转录组学、蛋白组学、代谢组学等生物多组学也取得了巨大进步，可以大规模低成本地获取参与信息传递的多维分子表型，包括 DNA、RNA 和蛋白质中元素的结构、修饰、功能和进化，以及它们之间的相互作用。海量生物组表型数据进一步促进了基于中心法则的信息传输和转换的细粒度剖析。对植物信息流的全方位研究对于基因组学基础研究和作物改良都有重大意义，例如研究识别与特定表型变异（人工诱变或自然变异）相关的基因组变异或者两者之间的因果关系。然而，分子表型中的丰富信息在很大程度上尚未得到有效探索，这使得从 DNA 序列到植物表型的端到端机制理解变得很困难。

随着深度学习与大数据技术的快速发展，开启了分子表型和植物表型研究的智能化研究时代。例如，通过深度学习的关联分析，可进行全转录组关联研究（TWAS），具有更短的信息传递路径和更少的信息转换步骤[2]。此外通过深度学习模型可以直接从上游分子表型或从基因组 DNA 序列预测分子表型[3]。本研究在概述深度学习概念方法的基础上，对近年来深度学习在分子表型建模与变异研究的应用场景和最新进展进行总结、概括和分析。同时，分析了深度学习方法在作物遗传改良中的应用，以期为相关研究人员提供参考。

2 深度学习：概念，方法与可解释性

2.1 深度学习基本概念

深度学习本质上是基于线性回归和一些激活函数的诸多分类器协同工作。深度学习中有很多神经节点，而不是传统统计学习中只有一个线性回归节点。在深度学习中，输入和输出之间有很多层。输入和输出之间的层称为隐藏层，节点称为隐藏节点。神经网络中的一个重要因素是受人类神经激发启发的激活函数，用于生成输入和输出之间的非线性关系。常用的激活函数例如 Sigmoid、Hyperbolic Tangent、ReLU。激活函数的作用是将数据转换和抽象成一个更可分类的平面。深度学习分类器需要借助梯度下降等数学工具来学习参数，尤其是在学习凸函数参数时效果显著。学习是通过最小化预测值和实际值之间的误差来完成的。本研究重点对深度学习的主流模型的架构和特征进行介绍，包括自动编码器、卷积神经网络、循环神经网络、生成对抗网络以及图神经网络等。

2.2 自动编码器（AE）

自动编码器（AutoEncoder）主要由编码器、解码器和隐藏层组成。自动编码器首先对输入信号进行编码，然后使用编码信号重建初始信号。该编码信号可以最小化初始信号和重构信号之间的误差。在编码和重构的过程中，编码器将输入数据映射到特定的特征空间。解码器将编码信号的特征映射回数据空间，然后重构初始数据。自动编码器的 3 个重要变体包括：稀疏自动编码器（Sparse Auto Encoder，SAE）、去噪自动编码器（Denoising Auto Encoder，DAE）和收缩自动编码器（Contractive Auto Encoder，CAE）。

2.3 卷积神经网络（CNN）

卷积神经网络具有共享权限的网络结构，可以有效降低网络模型的复杂度，同时也减少了权重的数量。处理高维图像效率更高，可以直接将图像作为整个网络的输入，有效避免传统算法复杂的特征提取和重构。

作为一个多层神经网络，卷积神经网络结构中的每一层由若干个二维平面组成，每个平面都有独立的神经元。卷积神经网络结构主要依靠共享权重、局部滑动窗口、下采样来保证输入数据的不变性。卷积神经网络的训练过程分为两个阶段。第一阶段是前向训练阶段，由 3 个步骤组成：根据给定的样本集随机选择样本；将样本作为初始数据放入网络；计算相应的输出数据。第二阶段是反向传播阶段，包括两个步骤：计算理想数据信息与输出数据信息的差值；根据反向传输的误差最小化方法调整权重矩阵。

2.4 循环神经网络（RNN）

与传统的神经网络不同，RNN 利用了网络中的序列信息。这一特性在许多应用中是至关重要的，包括 DNA 序列。在这些应用中，数据序列中的嵌入式结构传达了有用的知识。RNN 学习方式通过使用特定形式的存储器来模拟学习的知识随时间的动态变化，不仅分析当前的输入，而且对前序内容具备记忆能力。一个 RNN 可以被看作是短期记忆单元，包括输入层 x、隐藏（状态）层 s 和输出层 y，包括深度“输入到隐藏”“隐藏到输出”和“隐藏到隐藏”3 种模式。RNN 的一个主要问题是它对梯度消失和爆炸的敏感性。由于在训练过程中大量的小导数或大导数的乘法，梯度可能会衰减或爆炸。这种敏感性随着时间的推移而降低，意味着网络随着新输入的进入而忘记了最初的输入。因此，LSTM 被用来处理这个问题，在其递归连接中提供记忆块。每个记忆块包括存储网络时间状态的记忆单元，以及控制信息流的门控单元。

2.5 生成对抗网络（GAN）

生成对抗网络是基于博弈论的生成模型类。生成对抗网络没有明确地对数据分布进行建模，而是从中对样本进行建模。通过深度神经网络进行采样，神经网络将随机噪声作为输入，并将其转化为模型分布。生成对抗网络由两个神经网络组成：一种称为生成器；另一种称为鉴别器。这个模型被称为对抗模型，因为生成器不断地试图欺骗鉴别器，让其相信输入来自训练数据（真实数据），而鉴别器总是区分两者。这两个神经网络试图相互对抗。在获取这两个输入后，误差函数输出特定样本是真的还是假的概率，用于训练生成器和鉴别器的权重。

2.6 图深度学习（GNN）

深度学习或传统机器学习仅以向量的形式考虑欧几里得平面中的数据，例如图像、音频等。然而，图数据集具有以下 4 个不同特征，导致传统机器学习方法和深度学习方法在图数据领域应用的失效。

（1）不规则域图表示不规则域或非欧几里得数据，并不能像图像和音频一样，可以很容易地在欧几里得平。面或网格状结构中表示。导致许多数学运算不能直接应用于图数据。

（2）非静态结构。图可能具有不同的形状和结构，例如齐次、非齐次、有符号、无符号图等。图的细粒度可以以节点为中心（即链接预测、节点排名等）、或者以图为中心（例如图生成、图分类等）等。最常用的图表示方法是使用邻接矩阵。由于添加或删除节点后其形状会发生变化。

（3）可扩展性和并行化。图可能有数百万个节点和数十亿条边，庞大的数据成为传统深度学习模型训练的障碍，尤其是具有许多节点和隐藏层的模型。同时由于图中的每个节点都携带一些关于图中其他节点的信息，算法并行化也面临很多挑战。

（4）领域特定知识。在图上学习可能还需要了解领域特定知识。例如“药物 - 靶标”相互作用预测任务，其中药物化学分子结构可能有助于更好地预测。其他额外信息可能有助于将药物副作用作为特征进行预测。

图神经网络是一种输入为图数据而不是向量的神经网络。它学习表示每个节点的特征，进一步生成的特征可以用于任何与图相关的问题，例如节点分类、图分类、聚类等。每个节点的特征包含节点本身的特征与其邻居节点信息。

当前基于图神经网络，开发了许多衍生的深度学习模型，例如图卷积神经网络（GCN）和 GaphSage

等。图卷积神经网络分 3 步运行：卷积核、池化和 Flattening。

根据图神经网络中的不同核函数，可将图卷积神经网络分为两种类型：①空间方法。这类卷积运算不需要图的特征值。典型的工作包括 GAT 和 GaphSage。②谱方法。这类方法基于特征值，考虑了整个图结构以及各个图组件。

2.7　深度神经网络可解释性

可解释人工智能（Explainable AI，XAI）是以可理解的方式向人类解释，并呈现智能系统行为与决策的新一代人工智能。近年来，从模型内外 2 个角度对 XAI 模型的可解释问题提出了两大解决方案，包括“模型自身可解释”和“模型以外可解释”。前者是通过直接设计具有内在可解释性的算法实现模型的可解释功能，包括线性回归、逻辑回归在内的广义线性模型，以及梯度增强机、随机森林、极端梯度提升在内的树集成模型；后者将模型预测与解释分开，主要包括可视化解释、影响方法、基于实例的解释、基于知识的解释 4 种技术类型。

可视化解释是探寻深度神经网络等复杂模型内部工作机制最直接的途径，其技术方法主要包括：代理模型、部分依赖图（Partial Dependence Plot，PDP）和个体条件期望（Individual Conditional Expectation，ICE）。代理模型即用来解释复杂模型的简单模型，虽然计算量小，但其计算结果和高精度模型的计算分析结果相近。PDP 是一种图形表示，有助于可视化特定特征对机器学习模型预测结果的平均边际影响。ICE 是一种与 PDP 类似的图形表示，能深入到单个样本，分析某一特征变化对单个样本的影响，并给出每个样本的预测值。影响方法通过更改模型输入或内部参数来评估特征的重要性或相关性，并记录特征更改对模型性能的影响程度，以解释模型决策。影响方法主要有敏感性分析、层级相关性传播和特征重要性 3 种。敏感性分析通过使每个特征在可能的范围内变动来预测这些特征的变化对模型输出值的影响程度。层级相关性传播将模型决策的重要性信号从模型的输出层神经元逐层传播到模型的输入层，使模型的决策结果可在特征上找到解释，得到每个特征参与分类决策的贡献大小。特征重要性则是通过改变特征值，计算模型预测误差的变化，从而量化每个输入变量对模型预测结果的贡献。基于实例的解释技术通过选择数据集的特定实例来解释 AI 模型的行为，包括原型和批评解释，以及反事实解释。原型是指从数据集中选择的具有代表性的实例，数据集中的实例关系是由与原型的相似性决定的。为了避免过度泛化，数据集也需要展示批评点，即不能被一组原型有效代表的实例。模型可预测原型和批评的结果，以解释模型决策，并发现模型算法的弱点。反事实解释描述了一种因果关系，即“如果没有输入特征 *X*，则预测结果 *Y* 不会发生”，通过对原始实例的输入特征进行最小条件的更改，以获得不同预定输出结果的新实例，从而解释模型的决策行为。LIME 方法对模型进行局部可解释性分析。该方法通过扰动图片中的像素块观察模型预测性能的变化，如果模型预测性能下降，证明所删除的像素块是一个重要特征，实现对模型决策过程的解释。基于 Grad-CAM 的 CNN 可解释方法，利用加权梯度类激活映射显示出图像中对结果产生重要影响的区域。基于知识的解释主要包括提取内部知识和引入外部知识的解释方法。目前，基于知识提取的方法主要包括知识蒸馏和知识图谱。知识蒸馏是一种降低模型复杂度的模型压缩方法，可将信息从深层网络传递到浅层网络。

当主要目的不仅是准确预测而且是解释生物规则时，机器学习模型的可解释性和量化特征重要性对植物生物学研究来说变得至关重要。例如，在从植物基因组准确预测表型的同时，探究每个核苷酸的影响也非常重要。虽然深度学习提供了高精度的预测，但有时深度学习模型难以解释，这对于探索生物过程的推理至关重要。为了构建更多可解释的模型，SHAP（SHApley Additive exPlanations）为每个特征分配一个特定预测的重要性值。DeepLIFT（深度学习重要特征）分解神经网络对特定输入的输出预测，以定义重要特征。出于类似的目的，集成梯度旨在将深度网络的预测归因于其输入特征。另一方面，编码生物特征的选择在可解释性中也起着关键作用。最后，在运行模型

或解释结果之前，考虑测量错误或数据集提交过程中出现的错误也很重要。

3 深度神经网络在植物基因组学中应用的一般流程

根据数据标注情况可以分为两大类：有监督和无监督的深度神经网络。监督学习的目的是获得一个模型，将其预测变量（如 DNA 序列）映射到目标变量（如组蛋白标记）。目标变量可以是分类的（分类）或连续的（回归）。如果数据集中没有关于分类标签的标注，即为无监督学习，包括聚类和特征提取。

深度学习应用于基因组学的输入通常为将生物序列和分子表型分别作为预测变量和目标变量，其工作流程通常包括 4 个步骤，如图 1 所示。

（1）输入数据预处理。主要包括生物序列的检索和编码、分子表型的数字或分类表示，以及将预测“因子 - 目标”对正确拆分为训练、验证和测试集，通常采用生物序列之间的进化关系作为依据。

（2）模型构建和训练。主要包括模型架构和超参数的选择以及在训练集上训练模型。在训练期间应持续监控模型在验证集上的性能，以确定何时停止模型训练以避免欠拟合和过拟合。

（3）模型评估。评估训练模型在另一个数据集上的性能，称为测试集。用于衡量模型性能的指标取决于目标变量的性质：ROC 曲线下面积（auROC）常用于分类问题，R-squared 常用于回归问题。

（4）通过显著性或特征归因方法获取模型可解释性以识别生物序列中的功能元素。

4 深度学习在植物基因组学与作物育种方面的应用

4.1 深度学习与 DNA 和基因特性研究

深度学习已应用于大规模数据分析的多个领域，以解决基因组学、转录组学、蛋白质组学、代谢组学和系统生物学中的复杂生物学问题[4]。当前研究表明，DNA 形状在决定转录因子（TF）DNA 结合特异性方面起着重要作用[5]。深度学习模型可以使用大量数据类型，包括染色质可及性分析（例如 MNase-seq、DNase-seq、FAIRE）和其他基因组分析（例如微阵列、RNA-seq表达）。同样，对于转录因子 TF 结合，存在 ChIP-seq 数据、基因表达谱、DAP-seq（DNA 亲和纯化测序）

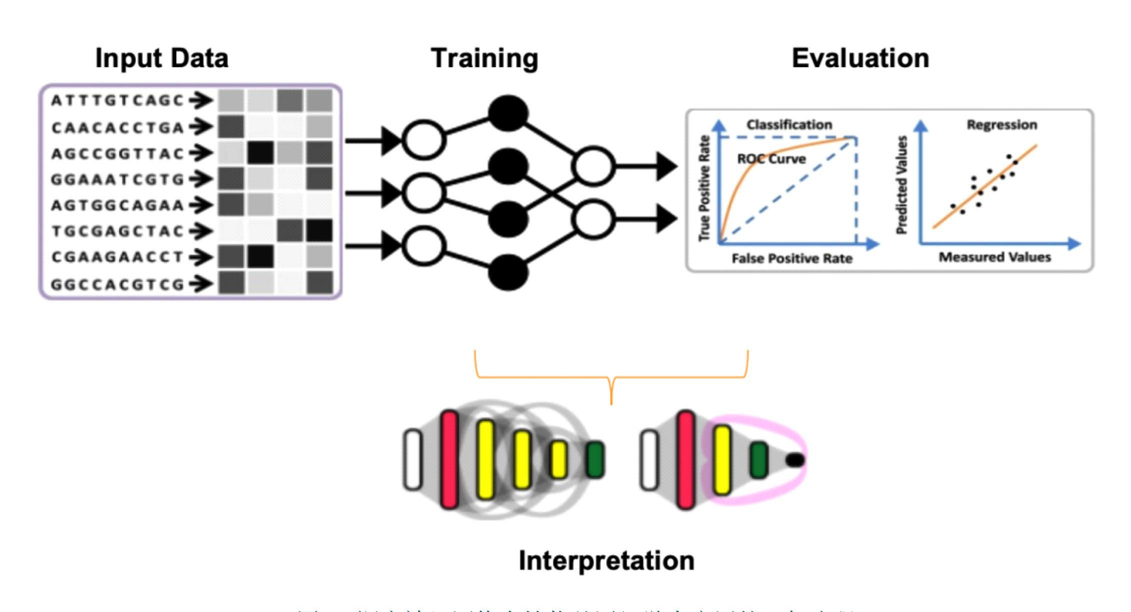


图 1　深度神经网络在植物基因组学中应用的一般流程

Fig.1 General process of deep neural network application in plant genomics

和 ampDAP-seq，通过使用扩增并去甲基化的 DNA 作为底物和组蛋白修饰来了解基因表达的潜在机制[6]。为了分析这些大规模数据集，当前有诸多深度学习方法来模拟 TF-DNA 结合特异性。为了预测 TF 结合特性，当前也有基于深度学习的方法。例如，了解 DNA 和 RNA 结合蛋白的序列特异性对于开发生物系统中的调控过程模型和识别致病变体至关重要[7]。

DeepBind[8]、DeepSEA[9]和 Basset[10]，是首批应用于基因组数据的卷积神经网络（CNN）。在 DeepBind 中，训练了多个单任务模型（参数的中位数为 1 586）来预测转录因子的体外和体内结合亲和力（即结合或未结合）和转录因子的体外结合亲和力。该方法始终比现有的非深度学习方法表现更好。DeepSEA 模型(52 843 119 个参数）从 DNA 元素百科全书（ENCODE）和 Roadmap Epigenomics 项目编译了 919 个 2.4M 非编码变体的染色质图，并测了 919 个染色质特征（人类 GRCh37 基因组）的存在与否，包括转录因子结合、DNA 可及性和给定 1 000BP 序列的组蛋白修饰。Basset（4 135 064 个参数）在给定 600bp 序列的情况下预测了 164 个二值化 DNA 可访问性特征。DeepBind 可以学习几个基序来预测 DNA 和 RNA 结合蛋白的结合位点。由 DeepBind 确定的特异性很容易被可视化为位置权重矩阵的加权组合或“突变图”，表明变异如何影响特定序列内的结合[3]。在 DeepSEA、DeFind[11]和 DFIM[12]中评估了功能性非编码变异的影响。DRNApred 用于区分 DNA 和 RNA 结合残基。由于数据集易于获得，上述这些方法通常是在组织或细胞系上进行训练和测试的。在玉米等具有大量重复元件和宽基因间区域的物种中，确定关键的基因组调控区域具有挑战性。为了应对这些挑战，基于自然语言处理的 k-mer 语法等方法已被用于以高效且精确地注释玉米品系中的调控区域[13]。使用大规模的 ChIP-seq 来重建玉米叶片中的网络，并训练机器学习模型来预测 TF 的结合和共定位。所得到的网络覆盖了 77%的表达基因，并显示出像现实世界网络一样的无标度拓扑结构和功能模块化。机器学习方法在模拟转录因子结合位点方面也发挥了重要作用。机器学习模型在植物生物学的几个方面已被证明是高效的，既可以单独或以组合方式从各种类型的测序数据中进行训练，还可以进一步整合其他信息，例如 DNase I 超敏数据，以更好地预测体内转录结合位点（TFBS）[13]。

总结而言，自最初应用以来，CNN 已被大量应用于基于 DNA 序列预测各种分子表型，并已成为新的最先进模型。应用包括分类转录因子结合位点[11]和预测分子表型，如染色质功能[14]，DNA 接触映射[15]，DNA 甲基化[16,17]，基因表达[18]，和 RBP 结合[19]。除了从序列中预测分子表型之外，CNN 还成功地应用于传统上由手工生物信息学方法解决的更多技术任务。例如，它们已被用于预测引导 RNA 的特异性[20]，增强的 Hi-C 数据分辨率[21]，从 DNA 序列预测起源的实验室和预测遗传变异体[22]。CNNs 也被用来模拟基因组中的长期依赖关系。尽管相互作用的调控元件可能位于未折叠线性 DNA 序列上的远处，但这些元件通常在实际的 3D 染色质构象中靠近。因此，从线性 DNA 序列建模分子表型，尽管是染色质的粗略近似，但可以通过允许长程依赖性并允许模型隐式学习 3D 组织的方面（例如“启动子 - 增强子”循环）来改进。在 Basenji[16]中，这是通过使用扩张卷积实现的，它启用了感受野达到 32KB。扩张卷积还允许使用 10KB 的感受野从序列中预测剪接位点。

在基因组学，RNNs 已被用于聚集细胞神经网络的输出用于预测单细胞 DNA 甲基化状态[17]，RBP 结合[23]，转录因子结合和 DNA 无障碍[24]。RNN 在 miRNA 生物学中也有应用：deepTarget[25]在从 mRNA-miRNA 序列对预测 miRNA 结合靶标方面比现有模型表现更好，并且 deepMiRGene[26]从 mRNA 序列及其预测的二级结构中比现有方法更好地预测前体 miRNA 的发生使用手工制作的功能。来自原始 DNA 测序数据的碱基调用是另一个应用 RNN 的预测任务。尽管 RNN 有诸多应用，但对于基因组学中常见的序列建模任务，缺乏对循环和卷积架构的系统比较。

4.2　深度学习在基因组学应用中的可解释性

在比较 CNN 和 k-mer 方法时，CNN 在特征提取

方面更有效。然而，CNN 通常被认为是黑匣子，因为对其输出的解释具有挑战性，并且可能涉及高计算成本。此外，他们的表现有多少来自于学习基本的生物规则，例如关键基序、基序关系和一般序列视角，这是相当不确定的。出于解释 DNA 的目的，k-mer 方法优于 CNN 和 RNN。使用 k-mers（或 k-tuples，k-gram）频率对序列进行分类是快速、准确、无参考和无对齐的。k-mer 是一种基于基因的方法，用于识别序列特征。通常，k-mer 频率向量与距离函数配对在一起，以测量任何一对序列之间的数量相似性。基于单词统计来恢复语义和句法线索很容易解释，但是，确定为什么以某种方式对序列进行分类并不像更传统的基于对齐的方法那样直接。然而，使用 k-mer 表示似乎是准确和快速分类的良好平衡。值得注意的是，也有结合 k-mer 方法和深度学习模型的例子[27]，尽管尚未系统评估这种方法对精度或可解释性的影响。

在线性模型等简单模型中，模型的参数通常衡量输入特征对预测的贡献。因此，在输入特征相对独立的情况下，可以直接用于模型解释。相比之下，深度神经网络的参数由于其冗余和与输出的非线性关系而难以解释。在复杂模型中，必须通过探测每个预测示例的“输入 - 输出”关系来间接得到特征重要性分数，也称为属性分数、相关性分数或贡献分数。特征重要性分数显示了给定输入中对模型预测最有影响的部分，从而有助于解释做出这种预测的原因。在 DNA 序列为基础的模型中，重要性分数可以表征序列基序，并因此广泛用于在基因组学[28]。特征重要性分数还可用于探测更复杂的上位相互作用[12]。

根据是使用输入扰动还是使用反向传播计算，特征重要性分数可以分为两大类。对于 DNA 序列为基础的模型中，诱导的扰动可以是单核苷酸取代或调节基序的插入。基于扰动的重要性得分的主要缺点是计算成本高，当需要计算整个数据集的重要性得分时，这一点就变得很明显。基于反向传播的特征重要性分数是更高效计算方式。在这些方法中，所有输入特征的重要性分数是使用通过网络的单个反向传播计算的，因此它们只需要两倍于单个预测的计算量。最简单的基于反向传播的重要性分数是 Saliency Maps[29]和 Input-Masked Gradients[30]。由于深度学习框架支持自动微分，这些分数可以在几行代码中有效地实现。

Saliency Maps、Input-Masked Gradients 或基于扰动的方法的一个问题是所谓的神经元饱和问题。为了解决此问题，提出了基于参考的方法，如 DeepLIFT 和 Integrated Gradients[31]。这些方法将输入特征与其“参考”值进行比较，从而避免饱和问题。在 DNA 序列的情况下，合理的参考值是原始序列的二核苷酸改组版本。我们注意到目前缺乏基因组学中特征重要性分数和不同参考值的严格基准。因此，建议尝试多种方法，并将它们与一些易于理解的示例或模拟数据进行比较。

最近提出了一种“可见神经网络”的方法，DCell 模型[32]，以提高内部神经网络激活的可解释性。DCell 对应于细胞内已知分子子系统的层次结构。神经网络中的节点对应分子子系统，例如信号通路或大蛋白质复合物，只有上游系统（例如小蛋白质复合物）是下游系统的一部分时，才允许两个节点（系统）之间的连接（如大的蛋白质复合物）。由于神经网络中的神经元对应已知概念，因此可以解释它们的激活和参数。这种方法仅适用于底层实体及其层次结构足够广为人知的任务，可能无法直接适用于实体或其层次结构通常未知的任务，例如转录因子结合。

4.3 图神经网络在基因组学中的应用

图结构数据，包括“蛋白质 - 蛋白质”相互作用网络和基因调控网络，在基因组学中无处不在。图表卷积神经（GCN）网络的使用的节点的各个特征中的曲线图和所述节点连接来解决图机器学习任务。GCN 依次应用多个图变换（层），其中每个图变换以非线性方式聚合来自相邻节点或边的特征，并用一组新特征表示节点或边。GCN 可以训练的任务包括节点分类，无监督节点嵌入（旨在找到节点的信息性低维表示），边缘分类和图分类。

GCN 已应用于许多生物和化学问题。例如，一种方法使用无监督的方法以无监督的方式从“蛋白质 - 蛋白质”相互作用网络中推导出蛋白质的新特征，然

后使用这些特征来预测不同组织中的蛋白质功能[33]。GCN 也被用于模拟多药副作用[34]。在化学中，曲线图的卷积已经成功地用于预测各种分子的性质，包括溶解性，药物功效和光电效率[35]。GCN 的基因组应用包括根据其他基因的表达[36]，研究了基因交互图（相同的路径、“蛋白质 - 蛋白质”、共同表达或研究论文文本关联）如何应用于深度模型，类似于图像上的卷积。探索了图卷积神经网络在基因组学的使用，通过结合基因嵌入以利用图信息。这种方法在低数据约束下为特定的任务提供了优势，但非常依赖于所用图形的质量。基因相互作用图的目的是捕捉基因之间的各种关系，并可用于创建更多的生物直观模型来进行机器学习。当前研究也试图通过利用这些图进行“单基因推断”（SGI）来评估这些图所提供的偏差。SGI 任务评估了与使用数据集中所有基因的基线相比，一个基因在特定图形中的邻居能多好地“解释”该基因本身。GCN为利用图的结构模式解决有监督和无监督的机器学习问题提供了有前景的工具，我们希望在未来看到更多的基因组学应用。

4.4 深度学习与基因组变异研究

给定生物序列作为预测因子，深度学习模型可用于预测分子表型（例如转录因子结合、表观遗传标记、染色质状态和基因表达水平）。深度学习模型最强大的部分是它们能够对新的、以前未见过的序列数据（即不在训练集中的数据）进行从头预测。

尽管自然种群中存在大量遗传变异，但可以对其中的一小部分进行深度学习模型训练，以预测所有其他变异（即整个变异空间）的影响。例如，在某些基因上训练的模型可用于对其他基因进行预测。这些不仅包括常见的等位基因，还包括低频和稀有变异，无论其影响程度如何。人类遗传学、精准医学和进化生物学的关键挑战包括破译基因表达的调控代码和理解基因组变异的转录效应。然而，由于非编码突变空间的巨大规模，这是极困难的。基于深度学习的框架 ExPecto 可以从 DNA 序列中准确预测突变的组织特异性转录效应，包括那些罕见的或未被观察到的突变。这使得探究基因表达的进化限制和突变疾病效应的初始预测成为可能，从而使 ExPecto 成为预测表达和疾病风险的端到端计算框架。ExPecto 是一个基于深度学习的框架，可以仅根据 DNA 序列预测突变的组织特异性转录效应。ExPecto 可以优先考虑 GWAS 位点的因果变体，并用于预测变体的疾病风险。由于在密切相关的物种中控制分子过程的生物学是保守的，在一个物种中训练的模型可以直接应用于密切相关的物种[37]。或者这些模型可以用作迁移学习中的教师模型密切相关物种的任务，促进知识从研究充分的物种（如拟南芥）迁移到相关但特征不佳的物种（如十字花科中的其他物种）。

提出了基于生物证据研究的自动建模（AMBER）[38]，这是一个完全自动化的框架，可以有效地设计和应用基因组序列的 CNN。AMBER 通过最先进的神经结构搜索（NAS）为用户指定的生物问题设计最佳模型。将 AMBER 应用于基因组调控特征的建模任务，并证明 AMBER 设计的模型的预测结果明显比同等基线的非 NAS 模型更准确，并匹配甚至超过了已发表的专家设计的模型。对 AMBER 架构搜索的解释揭示了其利用完整的计算操作空间来准确模拟基因组序列的设计原则。此外，说明了利用 AMBER 准确发现等位基因特异性结合和疾病遗传性富集的功能性基因组变异[38]。AMBER 为设计基因组学中准确的深度学习模型提供了一种有效的自动化方法。总之，深度学习模型可以极大地推动我们对终端表型的基因组变异的理解。

4.5 深度学习与蛋白质特性研究

4.5.1 CNN 与 RNN 在蛋白质预测中的应用

任何蛋白质的功能直接取决于其三级结构。蛋白质的三级结构可以通过综合分析各种蛋白质特性来揭示，例如二级结构、跨膜拓扑、信号肽、溶剂可及性、骨架二面角、无序到有序转变、接触图、模型质量、残基间接触、蛋白质相互作用位点、蛋白质紊乱和酶动力学。为了从头肽序列中提取重要的氨基酸特征，使用 CNN 方法开发了 DeepNovo[39]。基于串联质谱数据的新肽测序是猎枪蛋白质组学的关键技术，用于识

别没有任何数据库的肽和组装未知的蛋白质。然而，由于串联质谱的离子覆盖率较低，如果某些连续氨基酸的支持性片段离子全部丢失，则无法确定其顺序，这导致了从头测序的低精度。pNovo 3[40]使用一个学习排名框架来区分每个光谱的相似肽候选者。衡量每个实验光谱和其对应的理论光谱之间的相似性的 3 个指标被用作重要的特征，其中理论光谱可以由 pDeep 算法使用深度学习精确预测。在基于质谱的蛋白质组学中，多肽和蛋白质的鉴定和定量在很大程度上依赖于序列数据库搜索或光谱库匹配。由于缺乏准确的片段离子强度预测模型，降低了这些方法的实用性[41]。将 ProteomeTools 合成肽库扩展到 55 万条胰蛋白酶肽和 2 100 万条高质量串联质谱。并训练了一个深度神经网络 Prosit，在色谱保留时间和片段离子强度的预测方面取得显著提升。

蛋白质与蛋白质的相互作用（PPI）不断参与动态的病理和生物学研究过程中。因此，彻底理解 PPI 是非常重要的，有利于阐明疾病的发生，实现最佳的“药物 - 目标”治疗效果，并描述蛋白质的复合结构。

例如，文献使用图表示学习和结构特征的深度学习模型来预测 lncRNA 和蛋白质的相互作用，为了通用性和探索不同的模型设计原则，利用生物信息技术基于不同的特征提取和选择方法来开发 lncRNA- 蛋白相互作用预测算法,并基于互作关系进行功能推测。并在在拟南芥和玉米数据集上验证所提出方法的性能。

然而，与从不同物种和生物体获得的蛋白质序列相比，所揭示的“蛋白质 - 蛋白质”相互作用的数量相对有限。为了解决这一难题，许多研究工作都是为了促进发现新的 PPI 而进行的。在这些方法中，仅仅依靠蛋白质序列数据的 PPI 预测技术比其他需要广泛生物领域知识的方法更为广泛。为了预测二级结构，在深度学习模型中使用了相对溶剂可及性和残基间接触图 rawMS[42]。然而，深度学习算法在不同领域取得了成功，但由于覆盖率低和数据嘈杂，它们对 PPI 预测的有效性非常低。DPPI 成为一种能够从序列信息中预测 PPI 和同二聚体相互作用的新模型[43]。

提出了一种多模态的深度表征学习结构，将蛋白质的理化特征与来自 PPI 网络的图形拓扑特征相结合[44]。不仅考虑到了蛋白质序列信息，还考虑到了 PPI 网络中每个蛋白质节点的拓扑学表征。通过构建了一个堆叠的自动编码器架构，以及一个基于生成的元路径的连续词包（CBOW）模型来研究 PPI 预测。随后，利用监督下的深度神经网络来识别 PPI 并对蛋白质家族进行分类。8 个物种的 PPI 预测准确率从 96.76% 到 99.77% 不等，这是第一个用于研究 PPI 网络的多模态深度表示学习框架。

现有的“蛋白质 - 蛋白质”相互作用预测的计算方法大多集中在特征提取和特征组合上[45]。设计了一种名为 Res2vec 的新的残基表征方法来表示蛋白质序列。通过 Res2vec 得到的残基表征更精确地描述了原始序列的“残基 - 残基”相互作用，并为下游的深度学习模型提供了更有效的输入。结合有效的特征嵌入和强大的深度学习技术[45]，提供了一个通用的计算管道来推断“蛋白质 - 蛋白质”相互作用，即使是在蛋白质结构知识完全未知的情况下。

基于不同的蛋白质序列编码器，人们提出了大量的计算方法。一个蛋白质序列对的置信度分数可以被看作是对 PPI 的一种测量。一个蛋白质对的置信度分数越高，该蛋白质对就越可能发生相互作用。因此，引入了一个深度学习框架[46]，即序数回归和递归卷积神经网络（OR-RCNN）方法，从置信度的角度来预测 PPI。它主要包括两个部分：蛋白质序列对的编码部分和通过置信度分数预测 PPI 的部分。第一部分，应用两个具有共享参数的递归卷积神经网络（RCNN）构建两个蛋白质序列嵌入向量，可以自动从蛋白质对中提取稳健的局部特征和序列信息。在此基础上，通过元素相乘的方式将两个嵌入向量编码为一个新的嵌入向量。在第二部分中，通过考虑置信度分数背后的序数信息，使用序数回归来构建多个子分类器。多个子分类器的结果被汇总，得到最终的置信度分数。

4.5.2 ALPHAFOLD 在蛋白质预测中的应用

蛋白质对生命至关重要，了解其结构可以促进对其功能的机械性理解。通过巨大的实验努力，大约 10 万个独特的蛋白质的结构已被确定，但这只是数十亿

已知蛋白质序列中的一小部分。由于确定一个蛋白质结构需要数月至数年的艰苦努力，结构覆盖率成为当前研究的瓶颈。通过分析同源序列的共变性，可以推断出哪些氨基酸残基是接触的，这有助于预测蛋白质结构。AlphaFold 通过训练一个神经网络来对残基对之间的距离进行准确的预测，这比接触预测能传达更多的结构信息。利用这些信息，构建了一个能够准确描述蛋白质形状的平均力势。所得到的势可以通过一个简单的梯度下降算法进行优化，以生成结构，而不需要复杂的采样程序。即使对于同源序列较少的序列 AlphaFold 也能达到很高的准确性。AlphaFold 代表了蛋白质结构预测的一个相当大的进步[47]。

50 多年来，仅根据其氨基酸序列预测一个蛋白质采用的三维结构，即“蛋白质折叠问题”的结构预测部分，一直是一个重要的开放式研究问题。现有的方法远远达不到原子的准确性要求，特别是在没有同源结构的时候。AlphaFold2 提供了第一个可以定期预测蛋白质结构的计算方法，即使在没有类似结构的情况下也能达到原子精度。AlphaFold 的基础是一种新的机器学习方法，将有关蛋白质结构的物理和生物知识纳入深度学习算法的设计中，利用多序列排列的方式[48]。

AlphaFold2 通过结合新的神经网络架构 Evoformer 和基于蛋白质结构的进化、物理和几何约束的训练程序，大大提高了结构预测的准确性。提出了一个联合嵌入多序列排列（MSA）和成对特征的新架构，一个新的输出表示和相关损失，使准确的端到端结构预测成为可能，一个新的等价注意力架构，使用中间损失来实现预测的迭代完善，屏蔽 MSA 损失来与结构联合训练，使用自我蒸馏和自我估计准确性从无标签的蛋白质序列学习。Evoforme 是将蛋白质结构的预测视为三维空间中的图推理问题，其中图的边缘是由相近的残基定义。

BAEK 等探索了基于 DeepMind 框架的网络架构。他们使用了一个三轨网络来同时处理序列、距离和坐标信息，并取得了接近 DeepMind 的精度。通过 RoseTTA 折叠方法可以解决具有挑战性的 X 射线晶体学和低温电子显微镜建模问题，并产生准确的“蛋白质 - 蛋白质”复合物模型[49]。通过应用 AlphaFold2[50]，显著扩大了蛋白质组的结构覆盖范围，其规模几乎涵盖了整个人类蛋白质组（98.5%的人类蛋白质）。由此产生的数据集涵盖了 58%的残基，其中一个子集（占所有残基的 36%）具有非常高的置信度。同时在 AlphaFold 模型基础上开发了用于解释数据集的指标。AlphaFold2 从多序列排列（MSA）中编码的共同进化关系中预测蛋白质结构。尽管最近准确率大幅提高，但仍有 3 个挑战：①预测无法生成 MSA 的孤儿和快速进化的蛋白质；②快速探索设计的结构；③了解溶液中自发多肽折叠的规则[1]。提出了一个端到端的可区分的递归几何网络（RGN），能够在不使用 MSA 的情况下从单个蛋白质序列预测蛋白质结构。这个深度学习系统有两个新的元素：一个是蛋白质语言模型（AminoBERT），它使用转化器从数以百万计的未对齐的蛋白质中学习潜在的结构信息；另一个是几何模块，紧凑地表示 Cα 骨架几何。RGN2 在孤儿蛋白上的表现优于 AlphaFold2 和 RoseTTAFold（以及 trRosetta），并在设计序列上具有竞争力，同时在计算时间上实现了 106 倍的减少。

4.6　深度学习与作物育种研究

作物育种的一个重要组成部分是在环境适应和现代管理实践的背景下清除有害等位基因。过去 30 年，被概括为育种 3.0 时代，见证了标记辅助选择、关联分析和基因组预测的巨大胜利。值得注意的是，育种 3.0 时代标记辅助育种中使用的遗传变异不一定是农艺性状的因果变异。当育种者有能力大规模预测因果有益和有害变异时，可以通过编辑将有益等位基因直接引入优良种质，而不是通过在连锁位点携带有害等位基因的另一个供体亲本回交。同样，可以通过编辑有效地从基因组中清除有害等位基因。模拟研究表明，通过使用基因组编辑将有益的变异引入基因组，可以显着加速牲畜的育种。然而，由于基因型与环境之间的相互作用在作物物种中比在牲畜中更为突出，等位基因效应（无论是有害的、有益的还是适应性的）在作物物种中更具挑战性。理想情况下，特定于环境的模型或将环境因素作为额外输入的模型将缓解这个问题。

因此，可以合理地将深度学习模型预测的功能变异概念化为下一个育种时代的关键，即育种 4.0，其中作物物种的遗传改良在很大程度上取决于基因组编辑[4]。

在进行这种通过编辑繁殖的方法时，我们并不仅限于自然界中已知的有益变体。相反，我们享有完全的自由，可以根据我们的深度学习模型对感兴趣的生物过程的“理解”来创建新颖的有益等位基因。例如，编辑番茄 CLAVATA3 基因（SlCLV3）启动子[51]以增加果实大小并优化花序分枝[52]。由于 SlCLV3 启动子中缺乏功能注释，饱和启动子诱变采用 CRISPR/Cas9 系统，然后选择具有理想果实和花序特征的突变体。未来，通过从启动子序列预测基因表达水平的深度学习模型，可以通过单核苷酸分辨率的显着性评分识别 SlCLV3 启动子上的关键顺式元件，预测它们对 SlCLV3 基因的功能丧失影响表达，然后实施模型引导的启动子编辑。

创建具有特定功能的新基因组元素的另一种方法是在合成生物学中应用生成模型。例如，在学习现有启动子的突变空间后，可以训练模型以创建具有时空特异性的新启动子。然而，尽管变分自编码器和生成对抗网络等生成模型最近引起了广泛关注，但它们在合成生物学中的潜在应用仍然相当有限。一个例子是应用 GAN 来生成编码抗菌肽的合成 DNA 序列[53]。

4.7 无监督学习在基因组学及蛋白质特性中的应用

变异自动编码器（VAEs）和 GANs 是在深度学习领域出现的两种强大的生成方法。VAEs 是具有额外分布假设的自动编码器，使其能够生成新的随机样本。当前自动编码器已被用于填补缺失数据，提取基因表达特征，检测微阵列数据和大量 RNA，以寻找有意义的概率潜在表示[54]。自动编码器通常用于插补、降维和表征学习。因此，自编码器可以作为将映射从高维数据空间转换为低维特征空间的有效手段，从而提高聚类结果[55]。为了描述遗传对基因表达的影响，文献[56]建立了一个深度自动编码器模型来评估良好的遗传变异对基因表达变化的影响。文献[57]提出了 Adversarial Deconfounding AutoEncoder（AD-AE）方法去混淆基因表达潜在空间。通过联合训练网络生成嵌入，这些嵌入可以编码尽可能多的信息，而不会编码任何混杂信号。通过将 AD-AE 应用于两个不同的基因表达数据集，表明该模型可以：①生成不编码混杂信息的嵌入；②保存原始空间中存在的生物信号；③在不同的混杂域。

GANs 被认为是一种完全不同的生成模型的方法，它涉及两个神经网络，一个鉴别器和一个发生器网络。它们被联合训练，其中生成器旨在生成真实的数据点，而判别器则对给定样本是真实的还是由生成器生成的进行分类。GAN 已经被用来生成蛋白质编码的 DNA 序列[53]，并为蛋白质结合微阵列设计 DNA 探针。GANs 能够生成优于训练数据集中的序列，以更高的蛋白质结合亲和力来衡量[58]。在单细胞基因组学领域，GANs 已被用于模拟 scRNA-seq 数据和降维[59]。此外，作者通过扰动解释了 GANs 的内部表示。在 MAGAN143 中，作者使用一个由两个 GANs 组成的架构解决了来自不同领域的数据集，即 CyTOF 数据和 scRNA-seq 数据的对齐这一挑战性问题。使用生成模型来创建新的 DNA 元件、基因，甚至具有所需功能的调节回路，并将它们应用于作物改良将成为未来育种的发展重点之一。

5 总结与展望

本研究对近年来深度学习在植物基因组和作物育种研究领域的最新进展进行了总结梳理。总体来看，深度学习在基因组学研究诸多领域方向上取得了比传统方法更好的效果，深度学习在基因组学中的应用已经产生了具有科学和经济意义的早期应用。深度学习的优势主要体现在两个方面：①端到端学习，能够将多个预处理步骤整合到一个模型中；②多模态数据处理能力，可处理基因组学中极其异质的数据，包括序列、计数、质谱强度和图像。深度学习为基因组学与作物育种的研究拓展了全新的研究视角，随着算法精度不断提高，为促进表型与基因型组学的不同尺度关联研究带来新的机会。

深度学习当前已经在基因组学、转录组学、蛋白

质组学和合成生物学等领域取得诸多进展，可以为作物育种和植物基因组学领域提供强大驱动力，如完善基因组功能注释、挖掘新功能基因、预测植物表型、发现基因、RNA、蛋白质等物质的新分类模式，指导基因编辑。如何进一步将揭示与分子表型或终末性状相关遗传位点的关联作图与从 DNA 到分子表型信息流模型相结合，了解表型变异背后的因果变异，实现因果变异的优先级排序，提高表型预测准确性，进而加速遗传增益仍然是未来作物育种工作的巨大挑战。深度学习模型发展的巨大进步是分子表型预测，以及这些模型在通过连锁不平衡的计算机中断发现功能变异中的应用。研究用于全基因组识别有害和适应性变异的深度学习方法，是未来农业中基于编辑的作物遗传改良的先决条件。综上所述，深度学习为植物基因组学与作物育种的研究带来了巨大的机遇，为相关研究与应用提供新思路。深度学习模型可以极大地推动对终端表型的基因组变异的理解，并有希望应用于作物改良研究与实践中。

参考文献：

[1] CRICK F. Central dogma of molecular biology[J]. Nature, 1970, 227 (5258): 561–563.

[2] WAINBERG M, SINNOTT–ARMSTRONG N, MANCUSO N, et al. Opportunities and challenges for transcriptome–wide association studies[J]. Nature genetics, 2019, 51(4): 592–599.

[3] ERASLAN G, AVSEC Z, GAGNEUR J. Theis FJ deep learning: New computational modelling techniques for genomics[J]. Nature reviews genetics, 2019, 20(7): 389–403.

[4] XU C, JACKSON S A. Machine learning and complex biological data[J]. Genome biology, 2019, 20(1): 1–4.

[5] LAI X, STIGLIANI A, VACHON G, et al. Building transcription factor binding site models to understand gene regulation in plants[J]. Molecular plant, 2019, 12(6): 743–763.

[6] ZAMPIERI G, VIJAYAKUMAR S, YANESKE E, et al. Machine and deep learning meet genome–scale metabolic modeling[J]. PLoS computational biology, 2019, 15(7): E1007084.

[7] WANG H, CIMEN E, SINGH N, et al. Deep learning for plant genomics and crop improvement [J]. Current opinion in plant biology, 2020, 54: 34–41.

[8] DE LONG A, WEIRAUCH M T, et al. Predicting the sequence specificities of DNA–and RNA–binding proteins by deep learning[J]. Nature biotechnology, 2015, 33(8): 831–838.

[9] ZHOU J, TROYANSKAYA O G. Predicting effects of noncoding variants with deep learning –Based sequence model[J]. Nature methods, 2015, 12(10): 931–934.

[10] CHING T, HIMMELSTEIN D S, BEAULIEU–JONES B K, et al. Opportunities and obstacles for deep learning in biology and medicine[J]. Journal of the royal society interface, 2018, 15(141): 20170387.

[11] WANG M, TAI C, E W, et al. DeFine: Deep convolutional neural networks accurately quantify intensities of transcription factor –DNA binding and facilitate evaluation of functional non –coding variants[J]. Nucleic acids research, 2018, 46(11): E69–E69.

[12] GREENSIDE P, SHIMKO T, FORDYCE P, et al. Discovering epistatic feature interactions from neural network models of regulatory DNA sequences[J]. Bioinformatics, 2018, 34(17): i629–i637.

[13] QIN Q, FENG J. Imputation for transcription factor binding predictions based on deep learning[J]. PLoS computational biology, 2017, 13(2): E1005403

[14] KELLEY D R, RESHEF Y A, BILESCHI M, et al. Sequential regulatory activity prediction across chromosomes with convolutional neural networks[J]. Genome research, 2018, 28(5): 739–750.

[15] SCHREIBER J, LIBBRECHT M, BILMES J, et al. Nucleotide sequence and DNaseI sensitivity are predictive of 3D chromatin architecture[J]. BioRxiv, 2017: 103614.

[16] ZENG H, GIFFORD D K. Predicting the impact of non –coding variants on DNA methylation[J]. Nucleic acids research, 2017, 45 (11): E99.

[17] ANGERMUELLER C, LEE H J, REIK W, et al. DeepCpG: Accurate prediction of single –cell DNA methylation states using deep learning[J]. Genome biology, 2017, 18(1): 1–13.

[18] ZHOU J, THEESFELD C L, YAO K, et al. Deep learning sequence –Based AB initio prediction of variant effects on expression and disease risk[J]. Nature genetics, 2018, 50(8): 1171–1179.

[19] PAN X, SHEN H B. RNA –protein binding motifs mining with a

new hybrid deep learning based cross-domain knowledge integration approach[J]. BMC bioinformatics, 2017, 18(1): 1-14.

[20] KIM H K, MIN S, SONG M, et al. Deep learning improves prediction of CRISPR-Cpf1 guide RNA activity[J]. Nature biotechnology, 2018, 36(3): 239-241.

[21] ZHANG Y, AN L, XU J, et al. Enhancing Hi-C data resolution with deep convolutional neural network HiCPlus[J]. Nature communications, 2018, 9(1): 1-9.

[22] LUO R, SEDLAZECK F J, LAM T W, et al. Clairvoyante: A multi-task convolutional deep neural network for variant calling in single molecule sequencing[J]. BioRxiv, 2018: 310458.

[23] PAN X, RIJNBEEK P, YAN J, et al. Prediction of RNA-protein sequence and structure binding preferences using deep convolutional and recurrent neural networks[J]. BMC genomics, 2018, 19(1): 1-11.

[24] QUANG D, XIE X. DanQ: A hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences[J]. Nucleic acids research, 2016, 44(11): E107.

[25] LEE B, BAEK J, PARK S, et al. DeepTarget: End-to-end learning framework for microRNĂ target prediction using deep recurrent neural networks[C]. Proceedings of the 7th ACM international conference on bioinformatics, computational biology, and health informatics, 2016: 434-442.

[26] PARK S, MIN S, CHOI H, et al. DeepMiRGene: Deep neural network based precursor microrna prediction[J]. ArXiv preprint arxiv:1605.00017, 2016.

[27] SHEN Z, BAO W, HUANG D S. Recurrent neural network for predicting transcription factor binding sites[J]. Scientific reports, 2018, 8(1): 1-10.

[28] KELLEY D R, SNOEK J, RINN J L. Basset: Learning the regulatory code of the accessible genome with deep convolutional neural networks[J]. Genome research, 2016, 26(7): 990-999.

[29] SIMONYAN K, VEDALDI A, ZISSERMAN A. Deep inside convolutional networks: Visualising image classification models and saliency maps[J]. ArXiv preprint arxiv:1312.6034, 2013.

[30] SHRIKUMAR A, GREENSIDE P, SHCHERBINA A, et al. Not just a black box: Learning important features through propagating activation differences[J]. ArXiv preprint arxiv:1605.01713, 2016.

[31] SUNDARARAJAN M, TALY A, YAN Q. Axiomatic attribution for deep networks[C]. International conference on machine learning, PMLR, 2017: 3319-3328.

[32] MA J, YU M K, FONG S, et al. Using deep learning to model the hierarchical structure and function of a cell[J]. Nature methods, 2018, 15(4): 290-298.

[33] ZITNIK M, LESKOVEC J. Predicting multicellular function through multi-layer tissue networks[J]. Bioinformatics, 2017, 33(14): i190-i198.

[34] ZITNIK M, AGRAWAL M, LESKOVEC J. Modeling polypharmacy side effects with graph convolutional networks[J]. Bioinformatics, 2018, 34(13): i457-i466.

[35] KEARNES S, MCCLOSKEY K, BERNDL M, et al. Molecular graph convolutions: Moving beyond fingerprints[J]. Journal of computer-aided molecular design, 2016, 30(8): 595-608.

[36] DUTIL F, COHEN J P, WEISS M, et al. Towards gene expression convolutions using gene interaction graphs[J]. ArXiv preprint arxiv:1806.06975, 2018.

[37] KELLEY D R. Cross-species regulatory sequence activity prediction[J]. PLoS computational biology, 2020, 16(7): E1008050.

[38] ZHANG Z, PARK C Y, THEESFELD C L, et al. An automated framework for efficiently designing deep convolutional neural networks in genomics[J]. Nature machine intelligence, 2021, 3(5): 392-400.

[39] TRAN N H, ZHANG X, XIN L, et al. De novo peptide sequencing by deep learning[J]. Proceedings of the national academy of sciences, 2017, 114(31): 8247-8252.

[40] YANG H, CHI H, ZENG W F, et al. PNovo 3: Precise de novo peptide sequencing using a learning-to-rank framework[J]. Bioinformatics, 2019, 35(14): i183-i190.

[41] GESSULAT S, SCHMIDT T, ZOLG D P, et al. Prosit: Proteome-wide prediction of peptide tandem mass spectra by deep learning[J]. Nature methods, 2019, 16(6): 509-518.

[42] MIRABELLO C, WALLNER B. RawMSA: Proper deep learning makes protein sequence profiles and feature extraction obsolete[J]. Biorxiv, 2018: 394437.

[43] HASHEMIFAR S, NEYSHABUR B, KHAN A A, et al. Predicting

protein - Protein interactions through sequence-based deep learning[J]. Bioinformatics, 2018, 34(17): i802-i810.

[44] ZHANG D, KABUKA M. Multimodal deep representation learning for protein interaction identification and protein family classification[J]. BMC bioinformatics, 2019, 20(16): 1-14.

[45] LONGWELL S, SHIMKO T. Res2Vec: Amino acid vector embeddings from 3D-protein structure[J]. THRESHOLD, 30(22): 344.

[46] XU W, GAO Y, WANG Y, et al. Protein ⊨protein interaction prediction based on ordinal regression and recurrent convolutional neural networks[J]. BMC bioinformatics, 2021, 22(6): 1-21.

[47] SENIOR A W, EVANS R, JUMPER J, et al. Improved protein structure prediction using potentials from deep learning[J]. Nature, 2020, 577(7792): 706-710.

[48] JUMPER J, EVANS R, PRITZEL A, et al. Highly accurate protein structure prediction with AlphaFold[J]. Nature, 2021, 596(7873): 583-589.

[49] BAEK M, DIMAIO F, ANISHCHENKO I, et al. Accurate prediction of protein structures and interactions using a three-track neural network[J]. Science, 2021, 373(6557): 871-876.

[50] TUNYASUVUNAKOOL K, ADLER J, WU Z, et al. Highly accurate protein structure prediction for the human proteome[J]. Nature, 2021, 596(7873): 590-596.

[51] CHOWDHURY R, BOUATTA N, BISWAS S, et al. Single-sequence protein structure prediction using language models from deep learning[J]. BioRxiv, 2021.

[52] RODRíGUEZ-LEAL D, LEMMON Z H, MAN J, et al. Engineering quantitative trait variation for crop improvement by genome editing[J]. Cell, 2017, 171(2): 470-480, e8.

[53] GUPTA A, ZOU J. Feedback GAN (FBGAN) for DNA: A novel feedback-loop architecture for optimizing protein functions [J]. ArXiv preprint arxiv:1804.01694, 2018.

[54] LOPEZ R, REGIER J, COLE M B, et al. Deep generative modeling for single-cell transcriptomics[J]. Nature methods, 2018, 15(12): 1053-1058.

[55] M R, BEYAN O, ZAPPA A, et al. Deep learning-based clustering approaches for bioinformatics[J]. Briefings in bioinformatics, 2021, 22(1): 393-415.

[56] XIE R, WEN J, QUITADAMO A, et al. A deep auto-encoder model for gene expression prediction[J]. BMC genomics, 2017, 18(9): 39-49.

[57] DINCER A B, JANIZEK J D, LEE S I. Adversarial deconfounding autoencoder for learning robust gene expression embeddings [J]. Bioinformatics, 2020, 36(2): i573.

[58] KILLORAN N, LEE L J, DELONG A, et al. Generating and designing DNA with deep generative models[J]. ArXiv preprint arxiv:1712.06148, 2017.

[59] GHAHRAMANI A, WATT F M, LUSCOMBE N M. Generative adversarial networks simulate gene expression and predict perturbations in single cells[J]. BioRxiv, 2018: 262501.

Applications and Prospect Analysis of Deep Learning in Plant Genomics and Crop Breeding

HOU Xiangying[1], CUI Yunpeng[2*], LIU Juan[2]

(1. Zibo Academy of Agricultural Sciences, Zibo 255020; 2. Key Laboratory of Agricultural Big Data, Ministry of Agriculture and Rural Affairs, Institute of Agricultural Information, Chinese Academy of Agricultural Sciences, Beijing 100081)

Abstract: [Purpose/Significance] Advances in single-cell sequencing and high-throughput technology have made it possible for plant genomics to accumulate large quantities of data describing multidimensional genomic-wide molecular phenotypes at low cost. As powerful data mining tools, deep learning techniques can be utilized to further predict and interpret the acquired molecular phenotypes. In recent studies, deep learning has been shown to yield significant results in plant genomics and crop breeding research. However, a complete review of deep learning applications in plant genomics is lacking. [Method/Process] The input to deep learning applied to genomics is usually biological sequences and molecular phenotypes as predictor and target variables, respectively. We introduced the workflow from four views: input data pre-processing includes retrieval, coding, and splitting; model construction and training includes the selection of model architecture and hyperparameters; model evaluation and interpretability. Specifically, this paper introduces the background of deep learning approaches, including the latest graph neural networks; then it discusses two prominent issues in the intersection of genomics and deep learning with respect to gene characterization and protein characterization: 1) how to model the flow of information from plant genomic DNA sequences to molecular phenotypes; and 2) how deep learning models can be utilized to identify functional variation in natural populations? Specifically, the paper summarizes the current status of deep learning applications in related fields, which include deep learning and DNA and gene characterization research, interpretability of deep learning in genomics applications, graph neural networks in genomics, deep learning and genomic variation research, deep learning in protein prediction, ALPHAFOLD in protein prediction, deep learning and crop breeding research, and unsupervised learning in genomics and protein characterization. [Results/Conclusions] This article summarizes how traditional deep-learning algorithms, graph deep-learning, generative adversarial networks and interpretable AI are applied in current research in order to address these two problems. Finally, the prospects for deep learning in future plant genomics research and crop improvement are discussed. Overall, deep learning has provided better results than conventional methods in many genomics research directions, and the application of deep learning in genomics has yielded early applications of scientific and economic significance. Deep learning offers two distinct advantages: 1) end-to-end learning, with the ability to integrate multiple pre-processing steps into a single model; and 2) multimodal data processing capabilities that can handle extremely heterogeneous data in genomics. The advancement of deep learning has the potential to expand new research perspectives in genomics and crop breeding, and to facilitate larger-scale association studies in both phenotypic and genotypic genomics as algorithms become more accurate.

Keywords: plant genomics; crop breeding; deep learning; graph deep learning; review